UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

IN RE GOOGLE GENERATIVE AI COPYRIGHT LITIGATION

Case No. 23-cv-03440-EKL (SVK)

ORDER RE DISCOVERY DISPUTE

Re: Dkt. No. 302

Before the Court is the fourth discovery dispute in a line of disputes relating to Plaintiffs' efforts to identify the Named Plaintiffs' and putative class members' works that were used by Defendant Google LLC ("Google") in training its AI models. Dkt. 302; *see also* Dkts. 140, 159, 209. Plaintiffs first sought the provision of source code, which this Court denied. Based on Google's representations that it was the training datasets, not the source code, that would enable such identification, the Court approved (with some modifications) an inspection protocol for training datasets proffered by Google and agreed to by Plaintiffs. *See* Dkts. 140, 155. The subsequent disputes related to refinement/enforcement of this protocol. *See* Dkts. 159, 172, 208, 221. Plaintiffs now seek to compel Google, by whatever means possible, to "produce evidence sufficient to identify, or simply identify: (1) all Named Plaintiffs' copyrighted works in the datasets … used by Google to train the Models, and (2) all class member copyrighted works used for that purpose." Dkt. 302 at 6. The Court finds this matter suitable for resolution without oral argument. Civil L.R. 7-1(b). Having considered the Parties' submissions, the relevant law and the record in this action, the Court **DENIES** the request.

The Court denies Plaintiffs' requests for the same reasons it previously denied source code review: Google then represented and now maintains that "[t]he only record and the only means of determining what materials Google used to train its AI models is the training data itself." *See* Dkt.

302 at 7. Plaintiffs have proffered deposition testimony and other evidence suggesting that Google knows what works are acquired by its Core Data Acquisition Team ("CDA") and ingested through web crawling and that Google maintains records of books ingested into certain other data corpuses. Dkt. 302 at 4-6. But as Google points out, and as this Court has noted previously, none of this information "reveal[s] what survived the steps of preprocessing, filtering, deduplication, and construction of final training datasets, which are significantly different from what was included in [a] source corpus." Dkt. 302 at 9; Dkt. 280-1 at 2-3. Accordingly, consistent with its prior decisions, the Court is persuaded that only the training datasets will reveal what putative class works were ultimately used train Google's AI models.

Even if the Court were to order Google to "produce evidence sufficient to identify" the works used in the training datasets, Google's could only do so by having an expert work backward from the training datasets. Yet this is exactly what Plaintiffs acknowledge they have already done: "Plaintiffs submitted a methodology that reliably identifies Class Works in Google's training data." Dkt. 302 at 2. Indeed, Plaintiffs' true contention appears to be that they "should not have had to incur the time and expense of developing that methodology." *Id.* Plaintiffs' proposed relief, however, would not remedy this harm. Rather, it would have Google duplicate work that Plaintiffs admit was already done. "[A] party is not required to create a document where none exists." *Finjan, Inc. v. Juniper Network, Inc.*, No. 17-cv-05659-WHA (TSH), 2019 WL 2865942, at *1 (N.D. Cal. July 3, 2019) (cleaned up). Moreover, a defendant is generally not required to engage in analysis of records where the burden of deriving the answer from those records would be the same for the plaintiff. Fed. R. Civ. P. 33(d).

For all of the foregoing reasons, Plaintiff's request is redundant and not proportional to the needs of the case and is therefore **DENIED.**

SO ORDERED.

Dated: December 19, 2025

SUSAN VAN KEULEN
United States Magistrate Judge